Predicting Gender from Risk-Seeking Behaviors

Deepika Dilip, Rachel Tsong, and Adina Zhang

May 14, 2019

Introduction

One public health phenomenon that has been the subject of debate is the "Gender and Health Paradox", in which women experience decreased quality of life, yet have higher life expectancies than men. A theory that seeks to explain this pattern outlines increased risk-taking by men as a causal mechanism. By this logic, men would indicate more risk-taking behaviors than women when controlled for age and other confounders. We sought to test this theory by examining if risk-seeking behaviors could be predictive of gender.

Young People's Survey

For our analysis, we used a data set consisting of 1,010 responses from a 2013 survey administered among statistics students and their friends at Comenius University in Bratislava, Slovakia. This dataset is available to download from Kaggle. The survey consisted of 150 items including music and movie preferences, hobbies and interests, phobias, health habits, personality traits, views on life, and opinions, spending habits, and demographics. To test our hypotheses, we selected 19 variables that would demonstrate risk-seeking behavior.

After sectioning the data accordingly, we counted missing values to determine if that could affect our results. When only completed responses were considered (i.e. having values for every variable), we had a sample size of 942; 6.7% of the data was missing. As this value was less than 10%, we decided not to address this via imputation and opted for a deletion method (complete-case analysis). The limitations of this method will be addressed in the discussion.

Data Exploration

Method

Models

Logistic Regression

Discriminant Analysis

Linear

Quadratic

KNN

```r
knn.fit = train(x = dat1_train[,1:19],
                y = dat1_train$gender,
                method = "knn",
                preProcess = c("center", "scale"),
                tuneGrid = data.frame(k = seq(50, 120, by = 2)),
                trControl = ctrl)
```

SVM

Linear

Radial

```r
# Set tuning grid
radial_grid = expand.grid(C = exp(seq(-4, 5, len = 15)),
                          sigma = exp(seq(-8, -5, len = 5)))

set.seed(1)
# Fit radial SVM model
svm_radial = train(gender ~ .,
                   data = dat1_train,
                   method = "svmRadial",
                   preProcess = c("center", "scale"),
                   tuneGrid = radial_grid,
                   metric = "ROC",
                   trControl = ctrl)
```

Random Forest

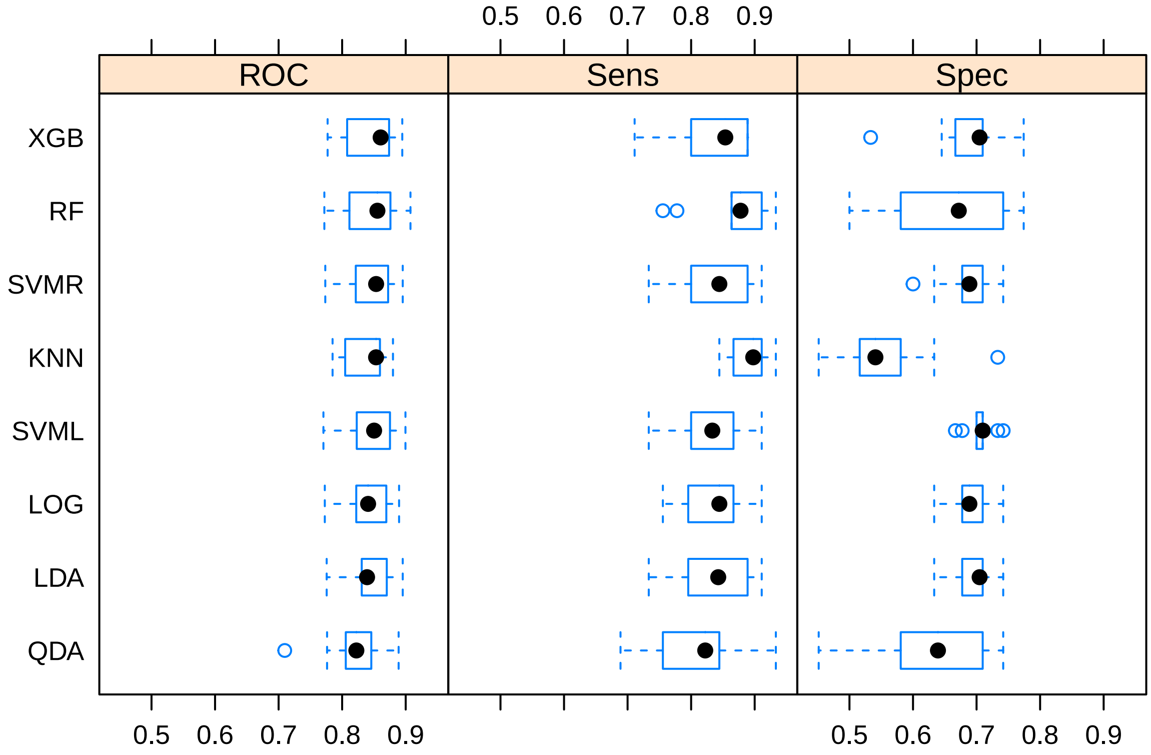
Extreme Gradient Boosting

```r
# Set tuning grid
xgbGrid = expand.grid(nrounds = seq(from = 50, to = 200, by = 50),
                      max_depth = c(2, 3, 4, 5, 6),
                      colsample_bytree = seq(0.5, 0.9, length.out = 5),
                      eta = 0.1,
                      gamma = 0,
                      min_child_weight = 1,
                      subsample = 1)

set.seed(1)
# Run boosting model using xgboost method
xgb.fit = train(gender~., dat1_train,
                trControl = ctrl,
                tuneGrid = xgbGrid,
                method = "xgbTree",
                metric = "ROC",
                importance = "impurity")
```

Model Selection

```r
resamp = resamples(list(LOG = log.fit, LDA = lda.fit,
                        QDA = qda.fit, KNN = knn.fit,
                        SVML = svm_linear, SVMR = svm_radial,
                        RF = rf.fit, XGB = xgb.fit))
bwplot(resamp)
```



Final Model

Conclusion

Appendix